DOCUMENT RESUME

ED 297 591 FL 017 500

AUTHOR Brutten, Sheila R.; And Others
TITLE Measuring the Instructional Sensitivity of ESL Reading Comprehension Items.
PUB DATE [88]
NOTE 20p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Comparative Analysis; *Difficulty Level; *English (Second Language); Higher Education; Language Tests; *Reading Achievement; *Reading Comprehension; Second Language Instruction; *Test Items; Test Theory

ABSTRACT

A study attempted to estimate the instructional sensitivity of items in three reading comprehension tests in English as a second language (ESL). Instructional sensitivity is a test-item construct defined as the tendency for a test item to vary in difficulty as a function of instruction. Similar tasks were given to readers at different proficiency levels and behavioral correlates were sought reflecting differences in time on task, as measured by the differences in content covered or "opportunity to learn." The assumption is that if behavioral differences can be attributed to differences in time on task, then those differences will be likely candidates for subsequent true-experimental design research in which they are systematically manipulated to ascertain whether instruction has been effective. The results were inconclusive and ran counter to expectations, suggesting the need for further research with a different variable and different instructional sensitivity indices. (Author/MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED297591

Measuring the Instructional Sensitivity of ESL Reading Comprehension Items

Sheila R. Brutten, John T. Mouw and
Kyle Perkins

Department of Linguistics
Southern Illinois University
Carbondale, IL 62901 U.S.A.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
K. Perkins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✔ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

(Measuring Instructional Sensitivity)

FL017500

Abstract

An analysis of variance procedure was used to estimate the instructional sensitivity of items in three English as a Second Language reading comprehension tests. The number of completed lessons was the factor investigated. No significant effect was noted in the results. Four reasons for the lack of significance are discussed: (1) an inadequate quantity of direct instruction; (2) the students' lack of linguistic readiness to accomodate instruction; (3) their lack of metacognition; and (4) their preoccupation with decoding. Academic engagement time and a set of instructional sensitivity indices are suggested for a future research program.

INTRODUCTION

This paper reports the results of a case study whose purpose was to estimate the instructional sensitivity of items in three English as a Second Language (ESL) reading comprehension tests. Instructional sensitivity is a test item construct which is supposed to reflect the role played by instruction in affecting examinee performances on individual test items. Haladyna and Roid (1981) defined instructional sensitivity as "the tendency for a test item to vary in difficulty as a function of instruction" (p. 40).

The basic paradigm employed in this research was to give similar tasks to readers in the form of a final comprehensive reading comprehension test and to look for behavioral correlates that reflected the differences in time on task as measured by the differences in content covered or "opportunity to learn" (Carroll, 1963). The assumption is that if behavioral differences can be attributed to differences in time on task, then those differences will be likely candidates for subsequent true-experimental design research in which those variables are systematically manipulated to ascertain whether or not instruction has been effective.

The hypothesis for this research was that the more time readers spend on a task, the better their performance on that task will be. Content covered has been shown to be a robust index of effective instruction, because it tends to be positively related to achievement gain (Anderson, Evertson, and Brophy, 1979; Barr, 1973-1974; Harris and Serwer, 1966; Rosenshine and Stevens, 1984).

Rosenshine and Stevens reviewed a number of correlational and experimental studies and reported correlations ranging from .2 to .7 between content covered and achievement gained. They noted that "it may be difficult to obtain higher correlations because there is not a perfect correspondence between the content given in a class and the content measured on the test" (p. 774).

METHOD

Subjects

This research was conducted at the Center for English as a Second Language at Southern Illinois University. Data were collected from three proficiency levels: 2, 3 and 4. At the Center, placement of students into the four full-time proficiency levels (1, 2, 3 and 4) is determined by the results of an Institutional TOEFL test which is administered to new students at the beginning of each term. Students who score 374 or lower are placed into level 1; students who score between 375 and 429 are placed into level 2; those who score between 430 and 469 are placed into level 3; and those who score between 470 and 524 are placed into level 4. If students score between 500 and 524, they are given additional testing (a composition and an oral interview) to determine whether they are proficient enough for full- or part-time university study or whether they need additional full- or part-time English study at the Center.

The distribution of subjects by native language was the following:

for level 2: Arabic, 7; Malaysian, 2; Spanish, 2; Thai, 1. For level 3: Arabic, 3; Spanish, 2; Malaysian, 4; Thai, 1; Farsi, 1; Japanese, 3; French, 1; Korean, 1; Senegalese, 1. For level 4: Arabic, 8; Spanish, 4; Malaysian, 4; Thai, 1; Farsi, 1; Japanese, 3; Korean, 3; Indonesian, 1; Chinese, 1.

Materials

All subjects were enrolled in eight-week intensive English classes. The reading classes met five days per week fifty minutes per day. Each proficiency level was broken into two sections. The level 2 sections used Lewis (1971) Reading for Adults as the text plus supplementary exercises prepared by the staff; level 3 sections used Zukowski/Faust, (1983) Between the Lines plus supplementary material; one section of level 4 used Long (1980) Reading English for Academic Study; the other section used Cowan (1982) Reading for College-Bound ESL Students.

During the term of instruction each instructor maintained a log of the number of instructional units completed for each reading skill. An instructional unit is defined for this research as a lesson or an exercise which provided instruction and practice on a particular reading skill. For each section of each level these data were coded and entered as a variable in the statistical analyses.

At the conclusion of the term subjects took a reading comprehension final examination. Each proficiency level had its own final examination. Each question was analyzed to determine what skill it was designed to assess.

Analysis

Because pre- and posttest data were not available for this research, we elected to conduct an analysis of variance study to estimate the instructional sensitivity of the items from the three different reading tests. The data from each proficiency level were analyzed separately. In the three separate analyses we were concerned with the following effects: (a) the number of completed lessons, to estimate instructional sensitivity; (b) skills, to estimate the differential difficulty of the various skills; and (c) the interaction of the number of completed lessons by the skills, to estimate differential instructional sensitivity across skills. We utilized the analysis of variance procedure because we had only postcomprehension item responses (i.e., no pre and posttest data) and because there is a robust tradition of using analysis of variance methods to estimate variability in items and scores (e.g., Haggard, 1958; Lewy, 1973; Rakow, Airasian, and Madaus, 1978; Wiley and Bock, 1967).

Design

As was mentioned in the introduction, this research activity was a case study; therefore, several aspects should be emphasized. First, the subjects were grouped into different proficiency levels according to a natural assembly process of TOEFL placement results (i.e., there was no randomization of class membership). Second, no systematic variation in the treatment factor (i.e., the number of lessons completed) was incorporated into the study because the texts and the entailed syllabi

used at the Center are set by the Center's curriculum committee, and the syllabi have a highly structured quality. We are aware that experimenters prototypically manipulate variables and observe changes in performance, but to preserve the ecological validity of the reading classroom this is neither desirable nor possible given that one's clientele are paying tuition and fees. In sum, many interesting problems in reading cannot be studied with true experimental designs for these reasons. For example, one cannot deliberately withhold instruction and practice on identifying the main idea of a passage from one section of a class while giving the other section twenty units of instruction and practice in that skill. And neither should one induce reading debility in order to study the effectiveness of various methods of remediation. Third, a constant environment was provided for all data elicitation because all tests were given during the last week of the term under timed examination conditions--constant, standardized testing procedures, therefore, reduced random variability or error.

RESULTS AND DISCUSSION

(Put Table 1 about here)

(Put Table 2 about here)

(Put Table 3 about here)

Table 1 presents the means, standard deviations, and standard

errors of measurement for the three tests; Table 2, the numbers of completed lessons for the groups; and Table 3, the analysis of variance summary table. Table 2 lists the number of completed lessons for the skills that were assessed by the tests. Other skills were taught in the three classes, but those skills were not assessed by the final examination; therefore, they were not entered in the Table 2 summary.

Table 3 indicates no significant effect for the number of completed lessons for proficiency level 2 ($p=.2471$) and for proficiency level 3 ($p=.5469$). For proficiency level 4 the factor approaches but does not quite reach significance ($p=.0547$). Also there was no significant interaction for the number of completed lessons by difficulty of skills: level 2 ($p=.5030$); level 3 ($p=.5957$); level 4 ($p=.4432$). In sum, the analysis of variance results indicated no significant effect for instruction.

There are several possible explanations for these insignificant results.

First, these students may not have had enough instruction in a given reading skill in order for that instruction to have made an impact on their acquisition or mastery of that skill, given that their interim second language competences were located at points on a continuum beginning with the mother tongue and ending with the target language. These students were exposed to a maximum of 33.33 hours of reading comprehension instruction (fifty minutes per day by five days by eight weeks). This is undoubtedly a liberal estimate given the fact that the actual engagement time was greatly reduced by the exchange of pleasantries, getting everyone settled,

administrative announcements, etc.

Second, given the developmental nature of these students' competences, they may not have been linguistically prepared to utilize the instruction they received in the most efficient fashion. Corder (1978) noted that it is only possible to teach someone new information, if that person is conceptually ready to accommodate it. In an earlier paper Corder (1967) distinguished input and intake:

> The simple fact of presenting a linguistic form to a learner does not neccessarily qualify it for the status of input, for the reason that input is "what goes in," not what is available for going in, and we may reasonably suppose that it is the learner who controls this input, or more properly his intake (p. 165).

Our results indicate that there was a substantial difference between what was available for "going in" and what "went in."

Third, these students may not have acquired the requisite metacognition which would allow them to recognize that their reading comprehension had failed on a particular skill and to select the necessary "fix-up" strategy to correct that deficiency. Brown, Armbruster, and Baker (1986) have defined metacognition as the "understanding of any cognitive process" (p. 49), and metacognition entails the learners' control of their own actions, their monitoring of their state of learning, their ability to plan strategies, and their knowledge of strategies for learning.

Fourth, since these students were still developing their reading competence, they probably tended to focus on reading as a decoding, "bottom-up" process and largely ignored reading as a meaning-getting,

"top-down," or encoding process. Reading is a complex, simultaneous interaction of bottom-up and top-down processing, and an overreliance on the bottom-up processing at the expense of the top-down processing probably has a negating influence on reading comprehension instruction.

DIRECTIONS FOR FUTURE RESEARCH

Although our results for this study were insignificant, and run counter to our intuitions and the current zeitgeist, we intend to freshen our endeavors to estimate instructional sensitivity by using a different variable and different instructional sensitivity indices in future studies. Our future research will utilize student engagement, an index of effective instruction. Fisher et al. (1978) noted that academically-engaged minutes of direct instruction in reading (systematic, structured instruction with immediate corrective feedback during student practice) is the most precise measure of student engagement. The academically-engaged minutes are calculated by multiplying the allocated time by the engagement rate. Thus, if the engagement rate was 53 percent, and the allocated time was 50 minutes per day, then the engaged minutes were 27 minutes per day. Rosenshine and Stevens (1984) reviewed the existing correlational studies on student engagement and found that the correlations between student engagement rate and reading gain ranged from .30 to .44. Student engagement will be used as a factor in our future analysis of variance studies.

Our future research will also utilize a set of item indices which

require pre- and post instruction data. There is an obvious, inherent danger in using the same test twice because students taking the test for a second time usually do better than those students taking it for the first time (Campbell and Stanley, 1966). However, that is a rival explanation for apparant gain that must be dealt with.

Some promising instructional sensitivity indices for reading research include (1) the pre-to-post difference index (PPDI) introduced by Cox and Vargas (1966); (2) the combined-samples (instructed and uninstructed examinees) point-biserial correlation (COMPBI) introduced by Haladyna (1974); (3) the normalized difference in item response theory item difficulty estimates for the uninstructed and instructed samples; and (4) three indices based on the Bayes' theorem (Bayes, 1763) which were proposed in Haladyna and Roid (1981).

PPDI is the difference in item difficulty (determined as the proportion of correct responses) observed when an item is presented to uninstructed examinees, and then after they have been instructed. According to Roid and Haladyna (1982), "this index is the simple difference between the difficulty observed in the post-instruction group and the difficulty observed in the pre-instruction group. It therefore ranges from -1.00 to +1.00, with indexes of zero or lower quite rare. In most instructional settings, this index typically ranges from .10 to .60" (pp.218-219).

The point-biserial correlation is a common procedure for computing item discriminability with point-biserial correlations of .25 and above being acceptable discrimination indices. Haladyna (1974) recommended combining uninstructed and instructed samples for the point-biserial

correlation, because more test-score variance is available with the combined samples, and COMPBI is less affected by range restriction. Haladyna noted that the COMPBI seems to function as a type of instructional sensitivity index.

The Bayes' theorem relates conditional and marginal probabilities. Haladyna and Roid (1981) proposed three instructional sensitivity indices based on the Bayes' theorem, and they require as input pre- and postinstruction data. The three Bayesian indices are: (a) B1, the probability that an examinee knows the content material given that the correct response was selected; (b) B2, the probability that an examinee does not know the content material given that the incorrect response was selected; and (c) B3, the probability that a correct decision will be made about the examinee's knowledge of the content given the results of performance on that item, that is, mastery or nonmastery. These instructional sensitivity indices have been successfully employed in an ESL context (Perkins, 1987).

CONCLUSION

Estimating the sensitivity of reading comprehension instruction is an important enterprise in our profession. The results permit us to investigate whether instruction has been faulty, whether the students have shown behavioral changes as a result of instruction, whether the apparatus and paradigm are robust, and whether the test items are flawed or inappropriate. Each of these propositions is worthy of its own research program.

REFERENCES

Anderson, R., Evertson, C., and Brophy, J. (1979) An experimental study of effective teaching in first grade groups. Elementary School Journal, 79, 132-222.

Barr, R. (1973-1974) Instructional pace differences and their effect on reading acquisition. Reading Research Quarterly, 9, 526-554.

Bayes, T. (1763) Essay toward solving a problem in the doctrine of chances. Philosophical Transactions, 53, 370-418.

Brown, A.L., Armbruster, B.B., and Baker, L. (1986) The role of metacognition in reading and studying. In J. Orasanu (ed) Reading Comprehension: From Research to Practice. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers, 49-75.

Campbell, D.T., and Stanley, J.C. (1966) Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally College Publishing Company.

Carroll, J.B. (1963) A model of school learning. Teachers College Record, 64, 723-732.

Corder, S.P. (1967) The significance of learner's errors. International Review of Applied Linguistics, 5, 161-169.

Corder, S.P. (1978) Language-learner language. In J.C. Richards (ed) Understanding Second and Foreign Language Learning. Issues and Approaches. Rowley, Massachusetts: Newbury House, 71-93.

Cox, R.C., and Vargas, J. (1966) A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the Annual Meeting of the Naional Council on Measurement in Education, Chicago, April.

Cowan, J.R. (1982) Reading for College-Bound ESL Students. Carbondale, Illinois: Center for English as a Second Language, Southern Illinois University.

Fisher, C.W., Filby, N.N., Marliave, R., Cahen, L.S., Dishaw, M.M., Moore, J.E., and Berliner, D.C. (1978) Teaching Behaviors, Academic Learning Time, and Student Achievement: Final Report of Phase III-B, Beginning Teacher Evaluation Study. San Francisco: Far West Educational Laboratory for Educational Research and Development.

Haggard, E.A. (1958) Intraclass Correlation and the Analysis of Variance. New York: Dryden Press.

Haladyna, T. (1974) Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 11, 93-100.

Haladyna, T., and Roid, G. (1981) The role of instructional sensitivity in the empirical review of criterion-referenced items. Journal of Educational Measurement, 18, 39-53.

Harris, A.J., and Serwer, B.L. (1966) The CRAFT project: Instructional time in reading research. Reading Research Quarterly, 2, 27-56.

Lewis, R.D. (1971) Reading for Adults. Book I. London: Longman.

Lewy, A. (1973) Discrimination among individuals v. discrimination among groups. Journal of Educational Measurement, 10, 19-24.

Long, M.H. (1980) Reading English for Academic Purposes. Rowley, Massachusetts: Newbury House.

Perkins, K. (1987) A comparison of instructional sensitivity indices. Unpublished paper.

Rakow, E.A., Airasian, P.W., and Madaus, G.F. (1978) Assessing school and program effectiveness: estimating teacher level effects. Journal of Educational Measurement, 15, 15-21.

Roid, G., and Haladyna, T. (1982) A Technology for Test-Item Writing. New York: Academic Press.

Rosenshine, B., and Stevens, R. (1984) Classroom instruction in reading. In P.D. Pearson (ed) Handbook of Reading Research. New York and London: Longman, 745-798.

Wiley, D. E., and Bock, R.D. (1967) Quasi-experimentation in educational settings: comment. School Review, 75, 353-366.

Zukowski/Faust, J. (1983) Between the Lines. New York: Holt, Rinehart and Winston.

Table 1

Means, standard deviations, standard errors of measurement, and estimates of reliability

Level	No. of subjects	No. of items	Mean	Standard deviation	Standard error of measurement	Estimate of reliability
2	12	25	19.98	4.19	1.78	.82
3	17	25	18.47	4.21	1.88	.80
4	25	25	15.68	4.37	2.14	.76

Table 2

The number of lessons completed

Proficiency level	Word meaning (in context)	Literal comp.	Inference (single string)	Inference (multiple strings)	Salients	Main idea
2A	0	14	3	3	0	0
2B	0	11	4	1	0	0
3A	6	1	0	2	0	6
3B	11	8	0	3	0	5
4A	7	3	0	5	6	2
4B	8	2	0	2	6	1

Table 3

Analysis of variance summary table

Proficiency level	df	F-value	p
Source of variation: difficulty of skills			
2	3	1.74	.1590
3	3	1.35	.2587
4	4	3.72	.0053*
Source of variation: number of completed lessons			
2	1	1.35	.2471
3	1	0.36	.5469
4	1	3.71	.0547
Source of variation: interaction of the number of completed lessons by the difficulty of skills			
2	2	0.69	.5030
3	2	0.52	.5957
4	3.	0.90	.4432

* $p < .01$